

Modeling Photographic Composition via Triangles

Zihan Zhou* · Siqiong He* · Jia Li · James Z. Wang

Received: date / Accepted: date

Abstract The capacity of automatically modeling photographic composition is valuable for many real-world machine vision applications such as digital photography, image retrieval, image understanding, and image aesthetics assessment. The triangle technique is among those indispensable composition methods on which professional photographers often rely. This paper proposes a system that can identify prominent triangle arrangements in two major categories of photographs: natural or urban scenes, and portraits. For the natural or urban scene pictures, the focus is on the effect of linear perspective. For portraits, we carefully examine the positioning of human subjects in a photo. We show that line analysis is highly advantageous for modeling composition in both categories. Based on the detected triangles, new mathematical descriptors for composition are formulated and used to retrieve similar images. Leveraging the rich source of high aesthetics photos online, similar approaches can potentially be incorporated in future smart cameras to enhance a person's photo composition skills.

Keywords Aesthetics · Photographic composition · Image segmentation · Triangle detection

Zihan Zhou* (✉) · Siqiong He* · James Z. Wang
College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
e-mail: zzhou@ist.psu.edu
(* = authors contributed equally)

Jia Li
Department of Statistics, The Pennsylvania State University,
University Park, PA, USA
e-mail: jiali@stat.psu.edu

Siqiong He
e-mail: hesiqiong@gmail.com

James Z. Wang
e-mail: jwang@ist.psu.edu

1 Introduction

With the rapid advancement of digital camera and mobile imaging technologies, we have witnessed a phenomenal increase of both professional and amateur photographs in the past decade. Large-scale social media companies, *e.g.*, Flickr, Snapchat, Instagram, and Facebook, further empowered their users with the capability to share photos with people all around the world. As over a billion new photos are added to the Internet daily, there is an increasing demand for creating machine vision application systems to manage, assess, and edit such content. As a result, *photo composition understanding* is becoming a noteworthy area that has attracted attention of the research community.

Composition is the art of positioning or organization of objects and visual elements (such as color, texture, shape, tone, and depth) within a photograph or a visual art work. Known principles of organization include balance, contrast, geometry, rhythm, perspective, illumination, and viewing path. Automated understanding of photo composition has been shown to benefit several applications such as summarization of photo collections and assessment of image aesthetics (Obrador et al., 2010). It can also be used to render feedback to the photographer on the aesthetics of her photos (Zhang et al., 2012; Yao et al., 2012), and to suggest improvements on the image composition through image re-targeting (Liu et al., 2010; Bhattacharya et al., 2010). In the literature, most work on image composition understanding has focused on image-based rules such as the simplicity of the scene, visual balance, the rule of thirds, and the use of diagonal lines. Because of their simplicity, these composition rules have been widely used to guide the photographers at the moment of their creative work.

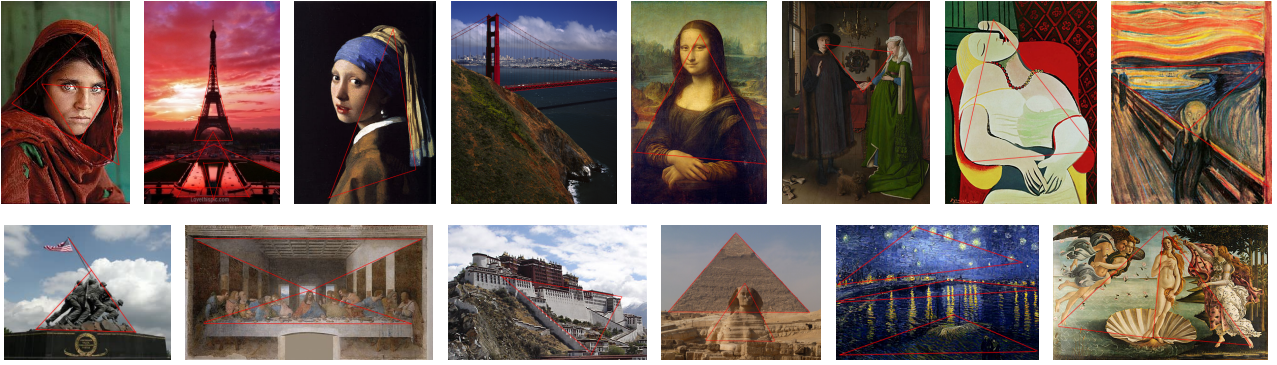


Fig. 1 The use of triangles in visual art and architectural works. The suggested triangles are indicated in red. From left to right: **(Row 1)** Afghan Girl, Eiffel Tower, Girl with a Pearl Earring, Golden Gate Bridge, Mona Lisa, The Arnolfini Portrait, The Scream. **(Row 2)** Iwo Jima Memorial, Last Supper, Potala Palace, Pyramid, Starry Night Over the Rhone, The Birth of Venus.

However, these rules are quite limited in capturing the wide variations in photographic composition. As an expansion, we hereby explore methods to identify an important composition technique, namely, *the triangle technique*. In pictorial art, good composition is considered as a congruity or agreement among the elements in a design (Lauer and Pentak, 2011). The design elements appear to belong together as if there are some implicit visual connections between them. One universal and interesting technique is to embed basic geometrical shapes in photographic compositions (Lauer and Pentak, 2011; Valenzuela, 2012). Human beings begin to learn about basic geometrical shapes such as circles, rectangles, and triangles at very young age. Moreover, because these shapes are instantly recognized, subjects bounded within such shapes or implicitly constructing such shapes are perceived as a unity. Among all basic geometric shapes, triangle is arguably the most popular shape utilized by professional photographers to make a composition more interesting. One can find numerous examples of the use of triangles in visual art and photographic works (Figure 1).

1.1 Category-sensitive Composition Modeling

We have developed category-sensitive approaches to detect the presence of the triangle composition in an image. The proposed methods can accurately locate a variety of triangles, even those which are carefully designed by professional photographers but difficult to be recognized by amateurs.

In our work, we focus on two major categories of photographs: natural or urban scenes, and portraits. When we contemplate most important photography genres, especially for consumers, arguably nature, travel, architecture, portrait, fashion, children, street, and people scenes are principal ones. With this work, we at-

tempt to show that the triangle technique can be applied to all of these main genres to improve the quality of the photo. A useful mobile app or a smart camera based on the triangle technique, for instance, can help the user/photographer with any of these photographic needs. The user can indicate to the system or device the particular type (roughly nature/urban vs. portrait/people) before taking the photo. Alternatively, the device can also determine the category automatically based on lens, focal length, the presence of face/people, GPS locations, among other available information. Other photography genres, such as animal, flower, macro, sport, and event photography, are frequently done by more sophisticated or professionally-trained photographers. The triangle technique is often not as important for them as some other techniques, *e.g.*, depth-of-field controlling, high speed telephoto, motion blurring, and emotion capturing. These genres are not covered in this work.

For both categories that we study here, “lines”, as a critical visual element in a picture, are exploited to detect triangles. From a technical viewpoint, *urban scenes* often have structures or objects that possess relatively clean straight lines. It is interesting to investigate whether a line-based technique can also be applied to more organic images such as *natural landscape photos* and *portraits*. In this paper, we propose to examine “contours” in the images, which can be considered as a generalization of lines in both natural scenes and portraits. By incorporating the contours in the line-based analysis, our work supports the usefulness of line-based techniques even on photos where straight lines are mostly absent. Because composition analysis is in nascence, we believe it is useful to show that the same line-based approach can be useful for both scenes with man-made objects and natural scenes or portraits.

While all of our techniques are based on the lines and contours, specific treatments for these two broad

categories of photos must be different. For example, with the nature/cityscape scenes, we can leverage information about the vanishing points in determining the triangle technique used, while we cannot make use of such information in most portraits. Despite the differences in algorithmic treatments, we believe it is desirable to have the two categories covered in the same publication because (1) it exemplifies the diversity of technical approaches needed to detect even simple visual elements like triangles in real-world scenarios, and (2) ultimately an application system using the triangle technique will likely incorporate all technologies described here under a uniform framework.

Finally, our methods can potentially benefit many composition-based applications. As an illustrative example, we apply the proposed methods for both categories to an image retrieval application which aims to provide amateur users with on-site feedback about the composition of their photos, in the same spirit as Yao et al. (2012). As we know, good composition highlights the object of interest in a photo and attract the viewer’s attention immediately. However, it typically requires years of practice and training for a photographer to master all the necessary skills on photo composition. An effective way for an amateur or an enthusiast to learn photography is through observing masterpieces, ideally with guidance, and establishing comprehension about photography. Professionally composed photographs can be valuable learning resources for beginners. Nowadays, thanks to the increased popularity of online photo sharing services such as Flickr and photo.net, one can easily access millions of photos taken by people all around the world. Such resources naturally provide us with opportunities to develop new and more effective ways for photographers to learn to improve composition skills.

1.2 System Overview

Figure 2 presents the framework of our composition analysis system. The user is first asked to indicate the type of photography he is currently engaged in (e.g., using a command dial selector, often available on consumer-level digital cameras). The system can provide assistance if the selected type is natural/urban scene or portrait. Future smartphone applications can also leverage other sources of information to automatically determine this.

For natural/urban scene, the user can take a test shot and upload it to our system as a query image. Given the query image, our system analyzes the overall composition, particularly the use of triangles, in the image. Then, it retrieves exemplar photos with simi-

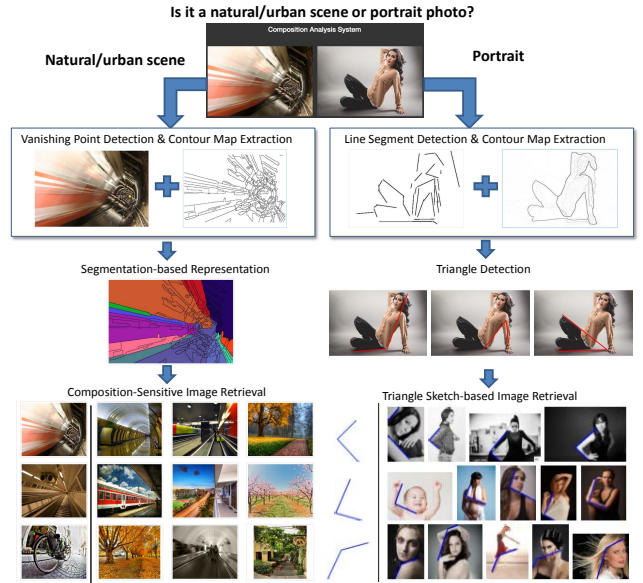


Fig. 2 Overview of our composition analysis system.

lar compositions from a collection of photos taken by experienced or accomplished photographers.

For portraits, the user can provide a sketch indicating the shape and orientation of the triangle he is looking for. Given the triangle sketch, our system will retrieve exemplar photos containing the specific triangular configuration from a collection of photos taken by experienced or accomplished photographers. Here, the use of triangle sketches is motivated by the following two facts. First, a professional portrait photo typically contains multiple triangles. The triangle sketch allows users of our system to examine one such triangle at a time, and gain a deep understand on such configuration across multiple images. Second, in practice, users such as magazine editors may wish to embed certain triangular configuration in the image in order to fill in a specific page layout.

For both categories, the retrieved exemplar photos can potentially serve as an informative guide for the users to achieve good compositions in their photos. In the following, we discuss the algorithms we have developed for each scene type, respectively.

1.2.1 Analyzing Triangles in Natural/Urban Scenes

In natural/urban scene photography, photographers often make use of the linear perspective effects in the images to emphasize the sense of 3D space in a 2D photo. According to the perspective camera geometry, all parallel lines in 3D converge to a single vanishing point in the image, generating a set of triangular regions. Figure 3 shows some examples. In order to convey a strong impression of 3D space and depth to viewers,

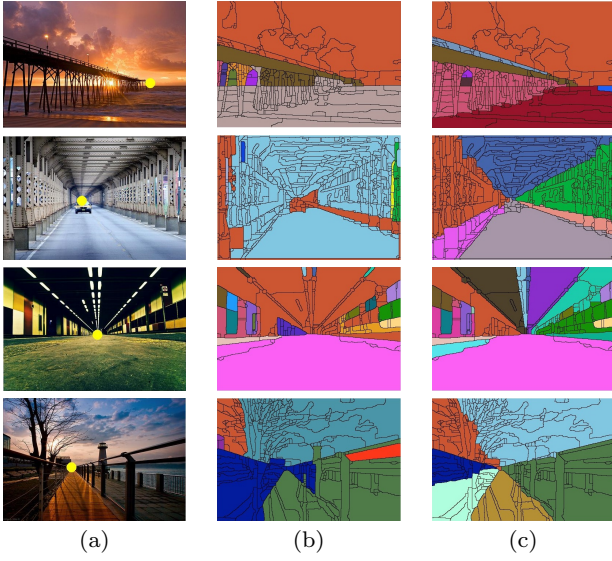


Fig. 3 Geometric image segmentation. (a) The original image with the dominant vanishing point detected by our method (shown as a yellow round dot). (b) Region segmentation map produced using a state-of-the-art method. (c) Geometric image segmentation map produced by our method.

the vanishing point has to lie within or near the image frame and associates with the dominant structures of the scene (*e.g.*, grounds, large walls, bridges). We regard such a vanishing point as the *dominant vanishing point* of the particular image. As one adjusts shooting angle, both the location of the dominant vanishing point as well as the sizes, shapes, and orientations of triangular regions relating to the vanishing point will change. An experienced photographer often utilizes such technique to produce various image compositions that convey different messages or impressions to viewers.

Accordingly, to model the composition for such scenes, we propose to *partition an image into photometrically and geometrically consistent regions according to the dominant vanishing point*. In our work, we assume that each geometric region can be roughly modeled by a flat surface, or a plane. Due to the perspective effect, these planes are naturally projected into *triangular regions* in the image. As shown in Figure 3(c), such a partition naturally provides us with a novel *holistic yet compact* representation of the 3D scene geometry that respects the perspective effects of the scene the image captured in, and allows us to derive a notion of relative depth and scale for the objects. Nevertheless, obtaining such a representation is a challenging problem for the following reasons.

First, given any two adjacent geometric regions in an image, there may not be a distinguishable boundary in terms of photometric cues (*e.g.*, color, texture) so that they can be separated. For example, the walls and

the ceiling in the second photo of Figure 3 share the same building material. Because existing segmentation algorithms primarily depend on the photometric cues to determine the distance between regions, they are often unable to separate these regions from each other (see Figure 3(b) for examples). To resolve this issue, we propose a novel hierarchical image segmentation algorithm that leverages significant geometric information about the dominant vanishing point in the image. Specifically, we compute a geometric distance between any two adjacent regions based on the similarity of the angles of the two regions in a polar coordinate system, with the dominant vanishing point being the pole. By combining the geometric cues with conventional photometric cues, our method is able to preserve essential geometric regions in the image.

Second, detecting the dominant vanishing point from an image itself is a nontrivial task. Typical vanishing point detection methods assume the presence of a large number of strong edges in the image. However, for many photos of outdoor scenes, such as the image of an arbitrary road, there may not be adequate clearly-delineated edges that converge to the vanishing point. In such cases, the detected vanishing points are often unreliable and sensitive to image noise. To overcome this difficulty, we observe that while it may be hard to detect the local edges in these images, it is possible to directly infer the location of the vanishing point by aggregating the aforementioned photometric and geometric cues over the entire image (Figures 3 and 8). Based on this observation, we develop a novel vanishing point detection method which does not rely on the existence of strong edges, hence works better for arbitrary images.

1.2.2 Modeling Composition in Portraits

Two fundamental questions are often raised when analyzing the composition of a portrait photograph: where are the human subjects in the photo and how do they pose? Traditional composition rules provide us with guidelines to answer the first question. For example, the rule of thirds suggests that putting the human subjects near the $1/3$ point of an image is more appealing than at the center. Based on these rules, several methods have been developed to model and assess the positioning of human subjects in a photo.

Nevertheless, the second question remains a challenge. To address this problem, we leverage an important observation in portrait photograph: *experienced portrait photographers often use triangle techniques to create interesting and good-looking poses for human subject*. For example, a widely-used rule for posing is that

one should try to avoid 90-degree body angles, because they often look unnatural and strained. In addition, triangle techniques are also frequently used to unify multiple human subjects and the surrounding environment, such as chairs and lamps, in the portrait photo.

Despite the popularity of triangle techniques in portrait photography, it is often difficult for less experienced amateurs to recognize such triangles, because most triangles do not have explicit edges and sometimes are even constructed by different objects. Moreover, triangles in portraits can be of various sizes, shapes, orientations, and appearances. Hence, our goal is to automatically detect potential triangles from professional photographers' work in order to help amateurs recognize and learn from the usage of triangle techniques.

Our algorithm can be divided into two steps: First, a line segment detection module is used to extract candidate line segments from an image, which are subsequently filtered using the global contour information in the image. Second, the filtered line segments are fed into a triangle detection module as the candidate sides of triangles. Specifically, a RANSAC algorithm is developed to randomly pick two sides from all the candidates and fit the triangle. Two metrics, *Continuity Ratio* and *Total Ratio*, are defined to evaluate the fitness of these triangles. Only those triangles with high fitness scores will be shown to the users.

1.3 Contributions

This paper makes the following main contributions:

- *Composition-sensitive retrieval for on-site feedback:* We developed triangle detection techniques for image composition understanding so that photographs with similar composition as the query photo can be retrieved from a collection of photos taken by professional photographers. User studies are conducted to verify to effectiveness of the retrieved exemplars in providing amateur users with useful information and guidance about photo composition.
- *Triangle detection and geometric image segmentation for natural/urban scenes:* We model the composition of typical natural/urban scene images by examining the perspective effects and partitioning the image into photometrically and geometrically consistent regions using a novel hierarchical image segmentation algorithm.
- *Dominant vanishing point detection:* By aggregating the photometric and geometric cues using our segmentation algorithm, we develop an effective method to detect the dominant vanishing point in an arbitrary image.

- *Triangle detection in portraits:* We propose a RANSAC algorithm to detect triangles in portraits with a variety of sizes, shapes, orientations, and appearances, with the goal of helping less experienced users to understand and learn from the usage of triangles in professional photographers' work.

Admittedly, our technique for natural/urban scenes and portraits cannot yet model all potential compositions in such photos, especially when there is a lack of triangles. While in this paper we focus on the use of triangle techniques in photography, we also point out that there are many works which study other important aspects of composition, including the semantic features (e.g., buildings, trees, roads) (Hoiem et al., 2005, 2007; Gould et al., 2009; Gupta et al., 2010). It would be ideal to integrate all these features in order to gain a deeper understanding of the image composition, but a thorough discussion on this topic is beyond the scope of this paper.

This article is an extension of our earlier work (Zhou et al., 2015). The primary new contributions are an expanded discussion on the use of triangles in photographic composition and a category-sensitive approach to study such usage (Section 1), a novel triangle detection method for portrait photos (Section 4), and its application to providing amateur users with on-site feedback about the composition of their photos (Section 6.2). We further demonstrate the effectiveness of our triangle detection method for portrait photos in Section 5.3.

2 Related Work

2.1 Composition Modeling

Standard composition rules such as the rule of thirds, golden ratio and low depth of field have played important roles in early works on image aesthetics assessment (Datta et al., 2006; Luo and Tang, 2008; Su et al., 2011). Obrador et al. (2010) later showed that by using only the composition features, one can achieve image aesthetic classification results that are comparable to the state-of-the-art. Recently, these rules have also been used to predict high-level attributes for image interestingness classification (Dhar et al., 2011), recommend suitable positions and poses in the scene for portrait photography (Zhang et al., 2012), and develop both automatic and interactive cropping and retargeting tools for image enhancement (Liu et al., 2010; Bhattacharya et al., 2010; Fang et al., 2014). Marchesotti et al. (2011) showed that, by aggregating statistics computed from low-level generic image features, one can achieve better

performance in assessing image aesthetics than using hand-crafted composition rules. In addition, Yao et al. (2012) proposed a composition-sensitive image retrieval method which classifies images into horizontal, vertical, diagonal, textured, and centered categories, and uses the classification result to retrieve exemplar images that have similar composition and visual characteristics as the query image. However, none of them study the use of triangles in photography.

2.2 3D Modeling and Segmentation of Images

Various methods have been proposed to extract 3D scene structures from a single image. The GIST descriptor (Oliva and Torralba, 2001) is among the first attempts to characterize the global arrangement of geometric structures using simple image features such as color, texture and gradients. Following this seminal work, a large number of supervised machine learning methods have been developed to infer approximate 3D structures or depth maps from the image using carefully designed models (Hoiem et al., 2005, 2007; Gould et al., 2009; Saxena et al., 2009; Nedovic et al., 2010) or grammars (Gupta et al., 2010; Han and Zhu, 2009). In addition, models tailored for specific scenarios have been studied, such as indoor scenes (Lee et al., 2009; Hedau et al., 2009, 2010) and urban scenes (Barinova et al., 2008). However, these works all make strong assumptions on the structure of the scene, hence the types of scene they can handle in practice are limited. Despite the above efforts, obtaining a good estimation of perspective in an arbitrary image remains an open problem.

Typical vanishing point detection algorithms are based on clustering edges in the image according to their orientations. (Kosecká and Zhang, 2002) proposed an Expectation Maximization (EM) approach to iteratively estimate the vanishing points and update the membership of all edges. Recently, a non-iterative method is developed to simultaneously detect multiple vanishing points in an image (Tardif, 2009). These methods assume that a large number of line segments are available for each cluster. To reduce the uncertainty in the detection results, a unified framework has been proposed to jointly optimize the detected line segments and vanishing points (Tretiak et al., 2012). For images of scenes that lack clear line segments or boundaries, specifically the unstructured roads, texture orientation cues of all the pixels are aggregated to detect the vanishing points (Rasmussen, 2004; Kong et al., 2009). But it is unclear how these methods can be extended to general images.

Image segmentation algorithms commonly operate on low-level image features such as color, edge, texture and the position of patches (Shi and Malik, 2000; Felzenszwalb and Huttenlocher, 2004; Li, 2011; Arbelaez et al., 2011; Mobahi et al., 2011). But it was shown in (Russell et al., 2009) that given an image, images sharing the same spatial composites can help with the unsupervised segmentation task.

2.3 Portrait Photo Analysis

Few studies have focused on aesthetic analysis of portrait images. Jin et al. (2010) studied the critical role of lighting in portrait photography. Varying lighting patterns shift lights and shadows on the face, change the area ratios between them, and generate 3D perception from the 2D photograph. While the learned artistic portrait lighting templates in that work are able to capture the arrangement of low level features, our work aims at modeling the usage of more holistic composition techniques in portrait photography. Zhang et al. (2012) developed a method to automatically recommend suitable positions and poses of people in natural scenes. However, the recommendation is based on matching simple 2D compositional features and manually labeled human body poses.

3 Natural/Urban Scene Composition Modeling

In this section, we present the technical details of our geometric image segmentation algorithm for triangle detection and composition modeling in natural/urban scene photos. Since our segmentation method follows the classic hierarchical segmentation framework, we give an overview of the framework and some of the state-of-the-art results in Section 3.1. In Section 3.2, we introduce our geometric distance measure for hierarchical image segmentation, assuming the location of the dominant vanishing point is known. The proposed geometric cue is combined with traditional photometric cues in Section 3.3 to obtain a holistic representation for composition modeling. In Section 3.4, we further show how the proposed distance measure, when aggregated over the entire image, can be used to detect the dominant vanishing point in an image.

3.1 Hierarchical Image Segmentation

Generally speaking, the segmentation method can be considered as a greedy graph-based region merging algorithm. Given an over-segmentation of the image, we

define a graph $\mathcal{G} = (\mathcal{R}, \mathcal{E}, W(\mathcal{E}))$, where each node corresponds to one region, and $\mathcal{R} = \{R_1, R_2, \dots\}$ is the set of all nodes. Further, $\mathcal{E} = \{e_{ij}\}$ is the set of all edges connecting adjacent regions, and the weights $W(\mathcal{E})$ are a measure of dissimilarity between regions. The algorithm proceeds by sorting the edges by their weights and iteratively merging the most similar regions until certain stopping criterion is met. Each iteration consists of three steps:

1. Select the edge with minimum weight:

$$e^* = \arg \min_{e_{ij} \in \mathcal{E}} W(e_{ij}) .$$

2. Let $R_1, R_2 \in \mathcal{R}$ be the regions linked by e^* . Set $\mathcal{R} \leftarrow \mathcal{R} \setminus \{R_1, R_2\} \cup \{R_1 \cup R_2\}$ and update the edge set \mathcal{E} accordingly.
3. Stop if the desired number of regions K is reached, or the minimum edge weight is above a threshold δ . Otherwise, update weights $W(\mathcal{E})$ and repeat.

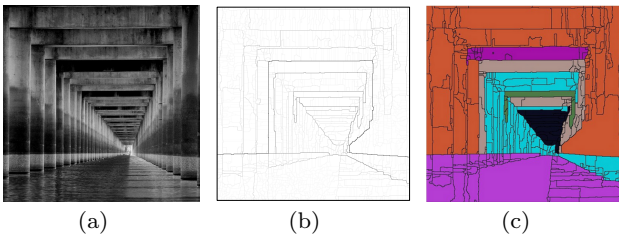


Fig. 4 Hierarchical image segmentation using photometric cues only. (a) The original image. (b) The ultrametric contour map (UCM) generated by Arbelaez et al. (2011). (c) The segmentation result obtained by thresholding the UCM at a fixed scale.

Various measures have been proposed to determine the distance between two regions, such as the difference between the intensity variance across the boundary and the variance within each region (Felzenszwalb and Huttenlocher, 2004), and the difference in coding lengths (Mobahi et al., 2011). Recently, Arbelaez *et al.* proposed a novel scheme for contour detection which integrates *global photometric information* into the grouping process via spectral clustering 2011. They have shown that this globalization scheme can help identify contours which are too weak to be detected using local cues. The detected contours are then converted into a set of initial regions (*i.e.*, an over-segmentation) for hierarchical image segmentation. We show an example of the segmentation result obtained by Arbelaez et al. (2011) in Figure 4. In particular, in Figure 4(b), we visualize the entire hierarchy of regions on an real-valued image called the ultrametric contour map (UCM) (Arbelaez, 2006), where each boundary is weighted by the

dissimilarity level at which it disappears. In Figure 4(c), we further show the regions obtained by thresholding the UCM at a fixed scale. It is clear that because the weights of the boundaries are computed only based on the photometric cues in Arbelaez et al. (2011), different geometric regions could be merged at early stages in the hierarchical segmentation process if they have similar appearances.

Motivated by this observation, we take the over-segmentation result generated by Arbelaez et al. (2011) (*i.e.*, by thresholding the UCM at a small scale 0.05) as the input to our algorithm, and develop a new distance measure between regions which takes both photometric and geometric information into consideration.

3.2 Geometric Distance Measure

We assume that a major portion of the scene can be approximated by a collection of 3D planes parallel to a dominant direction in the scene. The background, *e.g.*, the sky, can be treated as a plane at infinity. The dominant direction is characterized by a set of parallel lines in the 3D space which, when projected to the image, converge to the dominant vanishing point. Consequently, given the location of the dominant vanishing point, our goal is to segment an image so that each region can be roughly modeled by one plane in the scene. To achieve this goal, we need to formulate a dissimilarity measure which yields small values if the pair of adjacent regions belong to the same plane, and large values otherwise.

We note that any two planes that are parallel to the dominant direction must intersect at a line which passes through the dominant vanishing point in the image. Intuitively, this observation provides us with a natural way to identify adjacent regions that could potentially lie on different planes: If the boundary between two regions is parallel to the dominant direction (hence passes through the dominant vanishing point), these two regions are likely to lie on different planes. However, in the real world, many objects are not completely planar, hence there may not be a clear straight line that passes through the dominant vanishing point between them. As an example, if we focus our attention on the three adjacent regions R_1 , R_2 and R_3 in Figure 5, we notice that R_1 and R_3 belong to the vertical wall and R_2 belongs to the ceiling. However, the boundaries between the pair (R_1, R_2) and the pair (R_1, R_3) both lie on the same (vertical) line. As a result, it is impossible to differentiate these two pairs based on only the orientation of these boundaries.

To tackle this problem, we propose to look at the angle of each region from the dominant vanishing point

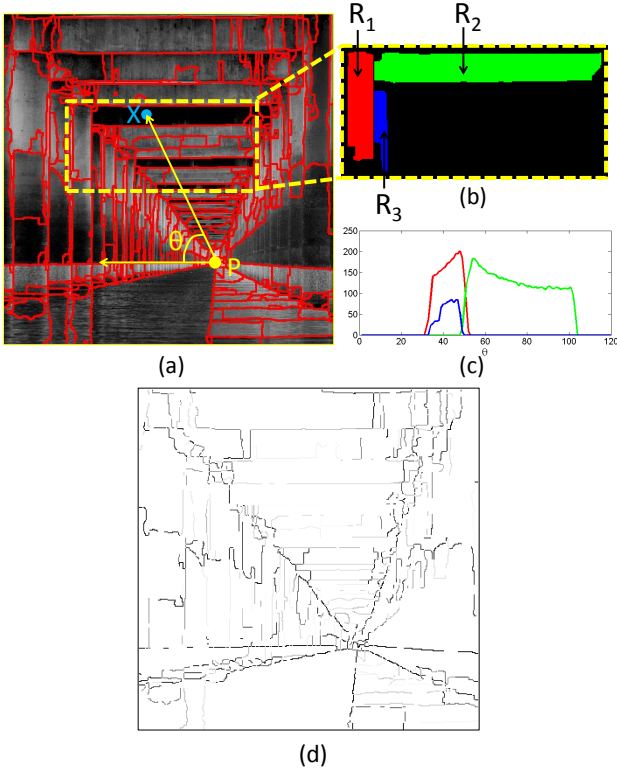


Fig. 5 Illustration of the computation of the geometric distance. (a) The over-segmentation map with the polar coordinate system. (b) Three adjacent regions from the image. (c) The histograms of angle values for the three regions. (d) The boundary map weighted by the geometric distance between adjacent regions.

in a polar coordinate system, instead of the orientation of each boundary pixel. Here, the angle of a region is represented by the distribution of angles of all the pixels in this region. Mathematically, let the dominant vanishing point P be the pole of the polar coordinate system, for each region R_i , we compute the histogram of the angle value $\theta(X)$ for all the pixels $X \in R_i$, as illustrated in Figure 5.

Let $c_i(\theta)$ be the number of the pixels in R_i that fall into the θ -th bin. We use 360 bins in our experiments. We say that one region R_i *dominates* another region R_j at angle θ if $c_i(\theta) \geq c_j(\theta)$. Our observation is that if one region R_i always dominates another region R_j at almost all angles, these two regions likely belong to the same plane. Meanwhile, if one region has larger number of pixels at some angles whereas the other region has larger number of pixels at some other angles, these two regions likely lie on different planes. This observation reflects the fact a plane converging to the vanishing point often divides along the direction perpendicular to the dominant direction because of architectural or natural structures, *e.g.*, columns and trees. Because perpendicular separation of regions has little effect on the

polar angles, the histograms of angles tend to overlap substantially.

Based on this observation, we define the geometric distance between any two regions R_i and R_j as follows:

$$W_g(e_{ij}) = 1 - \max \left(\frac{\sum_{\theta} \min(c_i(\theta), c_j(\theta))}{|R_i|}, \frac{\sum_{\theta} \min(c_i(\theta), c_j(\theta))}{|R_j|} \right),$$

where $|R_i|$ and $|R_j|$ are the total numbers of pixels in regions R_i and R_j , respectively. For example, as illustrated in Figure 5(c), R_1 dominates R_3 at all angles and hence we have $W_g(e_{1,3}) = 0$. Meanwhile, R_1 and R_2 dominate each other at different angles and their distributions have very small overlap. As a result, their geometric distance is large: $W_g(e_{1,2}) = 0.95$. In Figure 5(d), we show all the boundaries weighted by our geometric distance. As expected, the boundaries between two regions which lie on different planes tend to have higher weights than other ones. This suggests that, by comparing the angle distributions of two adjacent regions, we can obtain a more robust estimate of the boundary orientations than directly examining the orientations of boundary pixels.

Here, a reader may wonder why we don't simply normalize the histograms and use popular metrics like KL divergence or the earth mover's distance to compare two regions. While our intuition is indeed to compare the distributions of angles of two regions, we have found in practice that computing the normalized histograms could be highly unstable for small regions, especially at the early stages of the iterative merging process. Thus, in this paper we propose an alternative geometric distance measure which avoids normalizing the histograms, and favors large regions during the process.

3.3 Combining Photometric and Geometric Cues

While our geometric distance measure is designed to separate different geometric structures, *i.e.*, planes, in the scene, the traditional photometric cues often provide additional information about the composition of images. Because different geometric structures in the scene often have different colors or texture, the photometric boundaries often coincide with the geometric boundaries. On the other hand, in practice it may not always be possible to model all the structures in the scene by a set of planes parallel to the dominant direction. Recognizing the importance of such structures to the composition of the image due to their visual saliency, it is highly desirable to integrate the photometric and geometric cues in our segmentation framework to better model composition. In our work, we com-

bine the two cues by a linear combination:

$$W(e_{ij}) = \lambda W_g(e_{ij}) + (1 - \lambda) W_p(e_{ij}), \quad (1)$$

where $W_p(e_{ij})$ is the photometric distance between adjacent regions, and can be obtained from any conventional hierarchical image segmentation method. Here we adopt the contour map generated by Arbelaez et al. (2011).

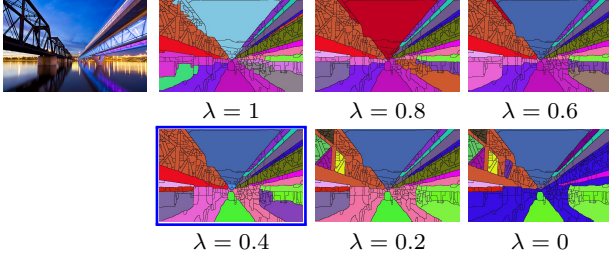


Fig. 6 Image segmentation results by integrating the photometric and geometric cues. Different weighting parameter λ have been used.

In Figure 6, we show the segmentation results of an image using our method with different choices of λ and a fixed number of regions K . Note that when $\lambda = 1$, only the geometric cues are used for segmentation; when $\lambda = 0$, the result is identical to that obtained by the conventional method (Arbelaez et al., 2011). It can be seen that using the geometric cues alone ($\lambda = 1$), we are able to identify most of the structures in the scene. Some of the boundaries between them may not be accurate enough (e.g., the boundary between the bridge on the left and the sky area). However, when $\lambda = 0$, the algorithm tends to merge regions from different structures early in the process if they have similar colors. By combining the two cues (e.g., $\lambda = 0.4$), we are able to eliminate the aforementioned problems and obtain satisfactory result. Additional results are provided in Figure 7. Our method typically achieves the best performance when λ is in the range of $[0.4, 0.6]$, as highlighted with blue boxes in Figures 6 and 7. We fix λ to 0.6 for the remaining experiments.

3.4 Enhancing Vanishing Point Detection

In the previous subsection we demonstrated how the knowledge about the dominant vanishing point in the scene can considerably improve the segmentation results. However, detecting the vanishing point in an arbitrary image itself is a challenging problem. Most existing methods assume that (1) region boundaries in the image provide important photometric cues about the location of the dominant vanishing point, and (2)

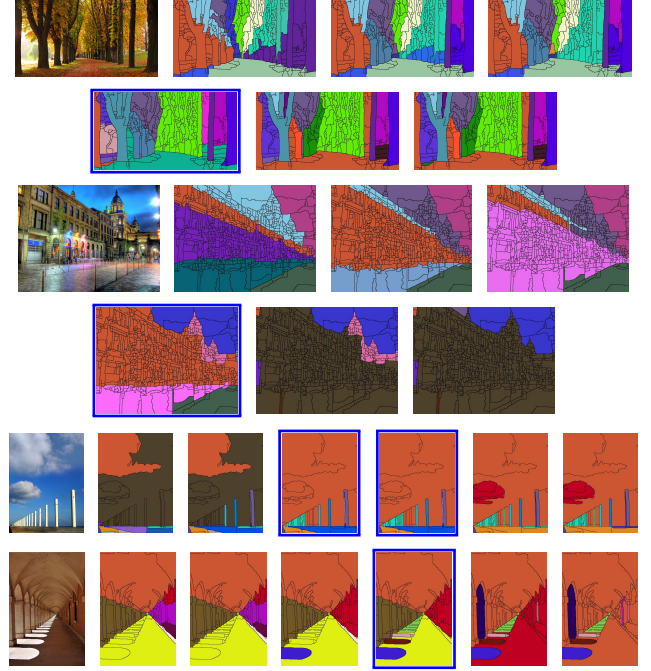


Fig. 7 Additional segmentation results. For each original image, we show the results in the order of $\lambda = 1, 0.8, 0.6, 0.4, 0.2$ and 0 .

these cues can be well captured by a large number of line segments in the image. In practice, we notice that while the first assumption is generally true, the second one often fails to hold, especially for images of natural outdoor scenes. This is illustrated in Figure 8: although human can easily infer the location of the dominant vanishing point from the orientations of the aggregated region boundaries, existing line segment detection algorithms may fail to identify these boundaries. For this reason, any vanishing point detection method relying on the detected line segments would also fail.

To alleviate this issue, we propose to use our geometric distance measure $W_g(e_{ij})$ to obtain a more robust estimation of the orientation of each boundary and subsequently develop a simple exhaustive search scheme to detect the dominant vanishing point. In particular, given a hypothesis of the dominant vanishing point location, we can obtain a set of boundaries which align well with the converging directions in the image by computing $W_g(e_{ij})$ for each pair of adjacent regions. These boundaries then form a “consensus set”. We compute a score for the hypothesis by summing up the strengths of the boundaries in the consensus set (Figure 8(c) and (d)). Finally, we keep the hypothesis with the highest score as the location of the dominant vanishing point (Figure 8(e)). Our algorithm can be summarized as follows:

1. Divide the image by an $m \times n$ uniform grid mesh.

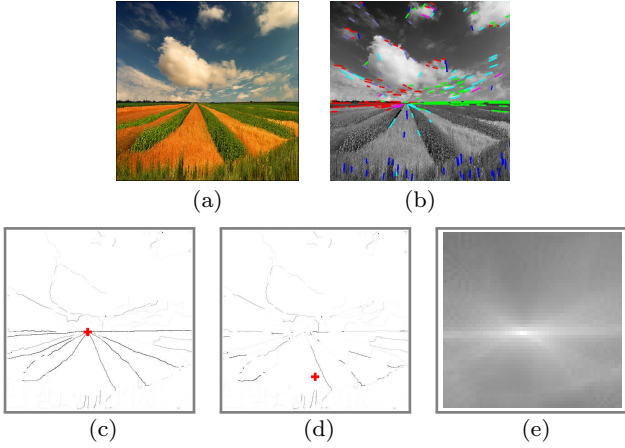


Fig. 8 Enhancing vanishing point detection. (a) Original image. (b) Line segment detected. (c) and (d) The weighted boundary map for two different hypotheses of the dominant vanishing point location. (e) The consensus score for all vertices on the grid.

2. For each vertex P_k on the grid, we compute the geometric distance $W_g(e_{ij})$ for all the boundaries in an over-segmentation of the image. The consensus score for P_k is defined as: $f(P_k) = \sum_{e_{ij} \in \mathcal{E}} W_p(e_{ij}) W_g(e_{ij})$.
3. Select the point with the highest score as the detected dominant vanishing point:
 $P^* = \arg \max f(P_k)$.

Here, the size of the grid may be chosen based on the desired precision for the location of the vanishing point. In practice, our algorithm can find the optimal location in about one minute on a 50×33 grid on a single CPU. We also note that the time may be reduced using a coarse-to-fine procedure.

In addition, we assume that the dominant vanishing point lies in the image frame because, as we noted before, only the vanishing point which lies within or near the frame conveys a strong sense of 3D space to the viewer. But our method can be easily extended to detect vanishing points outside the frame using a larger mesh grid.

4 Detecting and Modeling Triangles in Portrait Photography

In this section, we describe the algorithm for handling portrait photos. We first introduce the line segment detection algorithm (Section 4.1), and then discuss how triangles can be constructed from these line segments (Section 4.2).

4.1 Line Segment Detection

By examining high-quality portraits designed with the triangle technique (e.g., see Figures 16 and 17), we observe that triangles present in portrait photographs are often composed with parts of contours or edges, such as the contours of arms, body parts, wearing apparels, as well as edges formed by multiple human subjects. In general, a triangle is geometrically defined as a polygon with three corners and three sides, where the three sides are all straight line segments. However, contours of natural objects like humans or hats are often slightly curved. Therefore, to detect potential triangles in real images, our method needs to be able to identify such curved line segments, in addition to straight edges, in an image.

To this end, we employ the Line Segment Detector (LSD) proposed by von Gioi et al. (2010) to convert gradient map of an image to a set of line segments. It first calculates a level-line angle at each pixel to produce a *level-line field*. The level line is a straight line perpendicular to the gradient at each pixel. Then, the image is partitioned into *line-support regions* by grouping connected pixels that share the same angle up to a certain tolerance. Each line-support region is treated as a candidate line segment. Next, a hypothesis testing framework is developed to test each line segment candidate. The framework approximates each line-support region with a rectangle and compare the number of “aligned points” in each rectangle in the original image with the *expected* number of aligned points in a random image. A line segment is detected if the actual number of aligned points in a rectangle is significantly larger than the expected number.

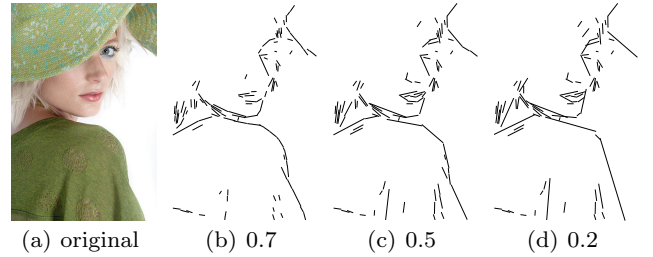


Fig. 9 LSD results with different density (as indicated).

A useful property of LSD is that, by approximating the line-support region using a rectangle of certain length, it is able to detect near-straight curves in the image. Further, it is easy to see that a larger rectangle with more unaligned points will be needed to cover a more curved line segment. Therefore, by setting a threshold on the *density* of a rectangle, which is defined

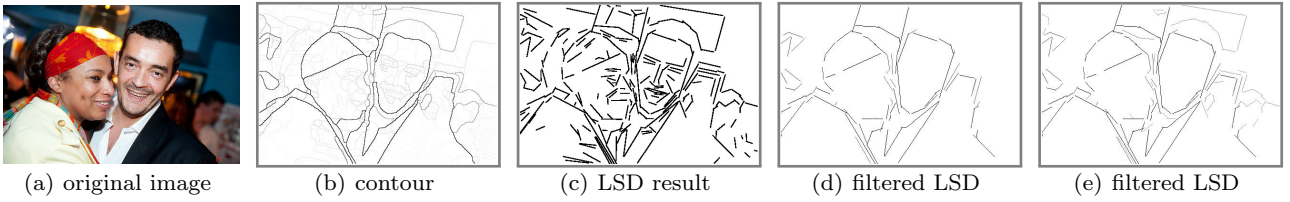


Fig. 10 Filtering the line segments. The parameter α is set to 0.5 and 0.7 in (d) and (e), respectively.

as the proportion of aligned points in the rectangle, we can control the degree up to which a curved line segment is considered. Figure 9 shows results of the line segment detector with different density values. In our experiments, the density is empirically set to 0.2.

While the line segment detector aims at extracting all potential line segments from an image using *local* image gradient cues, some line segments are more *globally* distinguishable, thus more visually attractive to viewers. To further identify such line segments, we combine the line segment detector with the ultrametric contour map (UCM) obtained by the same image segmentation algorithm (Arbelaez et al., 2011) we introduced in Section 3.1. Note that each pixel on the contour map holds a confidence level between 0 and 1, indicating the possibility of it being on a boundary. Given a line segment, we identify all the pixels falling in its support region and consider the maximum confidence level of all pixels as the confidence level of the line segment. Then, line segments with confidence levels under a certain threshold are removed, where the threshold is chosen based on the maximum confidence level present in an image. Specifically, assume the maximum confidence level of all the line segments in an image is C , where $C \in [0, 1]$, then we set the threshold as $(1 - \alpha)C$ and accept line segments whose confidence levels are within the range $[(1 - \alpha)C, C]$. The parameter α controls the number of accepted line segments. Smaller α filters out more line segments from an image, as shown in Figure 10. In this paper, we empirically set $\alpha = 0.5$.

4.2 Fitting Triangles

The line segment detector described above gives us a set of candidate triangle sides. Randomly selecting three non-parallel sides from the set generates a triangle. Thus, the problem of detecting a triangle can be converted into finding three non-parallel sides. However, although a triangle consists of three sides, we observe that a triangle can be uniquely determined as long as two sides are found, because the third side can be obtained by connecting the end points of the other two sides. Moreover, the presence of the third side is not as

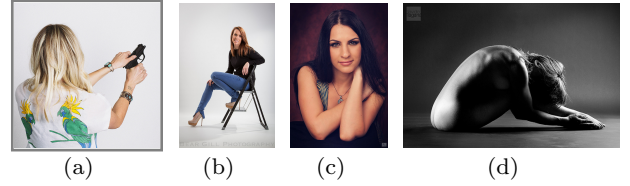


Fig. 11 Challenges in detecting triangles: Outliers and imperfect sides. (a) outliers: irrelevant objects, (b) outliers: multiple triangles, (c) imperfect sides: occlusion, (d) imperfect sides: bending curve.

important in the practical usage of triangle technique because viewers can easily “complete” the geometric shapes themselves. As a result, our problem is reduced to fitting a triangle using two non-parallel line segments selected from the candidate set.

Two major challenges still exist in fitting triangles using the extracted line segments: (1) there is a large number of irrelevant line segments; and (2) the sides of a triangle are imperfect in real images. For example, as shown in Figure 11(a), some objects in a photograph are irrelevant to the use of triangle technique, even though they have high-contrast contours (*e.g.*, the bird pattern on the woman’s shirt). The line segments produced by such objects are all outliers. In addition, multiple triangles often exist in one image (*e.g.*, Figure 11(b)). Thus, line segments from different triangles should also be considered as outliers w.r.t. each other. In Figures 11(c) and 11(d), we further show some examples of imperfect triangle sides. As shown, these sides may be broken into several parts because of occlusion or artificial effect introduced by the line segment detector. For instance, the girl in Figure 11(c) has her right arm occluded by her left arm. As a result, the contour of her right arm is broken into two line segments. In Figure 11(d), the contour of the back of the human subject is too curved to be approximated by a single straight line. Therefore, the line segment detector approximates it using two straight line segments with slightly different orientations.

In order to tackle these two challenges, we employ a modified RANSAC (RANDOM Sample Consensus) algorithm in favor of its insensitivity to outliers. RANSAC

is an iterative method that robustly fits a set of observed data points (including outliers) to a pre-defined model. Our algorithm includes three steps:

1. Two non-parallel line segments are randomly selected from the candidate set and extended to generate two lines on which the two triangle sides lie.
2. All the candidate line segments within neighborhoods of the two lines are projected onto the correspond lines, resulting in a number of projected pixels. Triangle sides are then constructed from all the projected pixels.
3. Once two triangle sides are constructed, two metrics *Continuity Ratio* and *Total Ratio* are calculated to measure the fitness and significance of the triangle, respectively. Triangles with high scores are accepted.

In the remainder of this subsection, we describe each step in details.

4.2.1 Identifying Sides From Line Segments

By extending the two randomly selected line segments to two lines, we obtain the shared end point of the two sides, *i.e.*, the intersection of the lines. Moreover, two intersecting lines generate four different angles with four different opening directions: upwards, downwards, leftwards, and rightwards. Each angle corresponds to a category of triangles that contain this angle and two sides of varied lengths. Given one of the four angles, once the lengths of its two sides are determined, a unique triangle can be constructed.

4.2.2 Fitting All Segments on Sides

In this step, we first mark all line segments within neighborhood of the two straight lines formed in the previous step as inliers and those falling outside neighborhood as outliers. The neighborhood region of a straight line $l : ax + by + c = 0$ is defined to be

$$N(l) = \{(x, y) \in \mathbb{R}^2 \text{ and } \frac{|ax + by + c|}{\sqrt{a^2 + b^2}} \leq d_{nb}\},$$

i.e., the group of pixels whose distances to the straight line are smaller than a certain threshold d_{nb} . In this paper, we fix $d_{nb} = 5$ pixels. Then, all the inlier line segments with respect to line l can be calculated as $I(l) = S \cap N(l)$ where S is the set of all candidate line segments. Note that, if a line segment is cut into two parts by the neighborhood boundary, the part of line segment falling within the neighborhood is included as an inlier, whereas the other part is considered as an outlier. In Figure 12(c), we illustrate how the neighborhood of a straight line is utilized to partition all the line

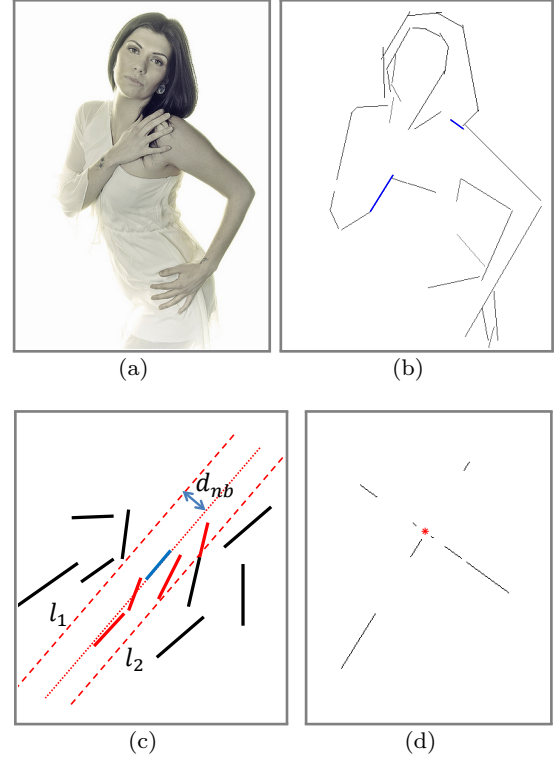


Fig. 12 Illustration of the RANSAC algorithm. (a) original image, (b) candidate sides, (c) neighborhood region, and (d) fitted pixels.

segments into inliers and outliers. The blue line segment is selected from the candidate line segment set and extended to a straight line l . Lines l_1 and l_2 designate boundaries of the neighborhood region. Both of them have a distance d_{nb} from l . The red line segments and black line segments represent the inliers and outliers, respectively.

Next, all the pixels on the inlier line segments are projected onto the straight line. A pixel on the straight line is called a *projected pixel* if there is at least one pixel on any inlier line segment that is projected to this pixel. We denote the set of all projected pixels as

$$P(l) = \{(x', y') \in l \mid \exists (x, y) \in I(l), (x - x', y - y') \perp l\}.$$

Figure 12 illustrates the fitting process. Figure 12(b) shows all the line segments extracted from the image shown in Figure 12(a). The two line segments colored in blue are randomly selected from the candidate set. Subsequently, two straight lines, l and \tilde{l} , are generated by expanding the two line segments and their neighborhood regions are identified. Finally, inliers within their neighborhoods are projected onto the straight lines, as shown in Figure 12(d).

Here, we note that the projected pixels typically scatter along the entire straight line. To form a triangle,

we divide the line into two *half lines* at the intersection point. Formally, a half line here is defined as a straight line extending from the intersection point indefinitely in one direction only. It is easy to see that, for each pair of straight lines l and \tilde{l} , four triangles can be formed using different pairs of half lines. Therefore, when evaluating the fit of a triangle, we only consider the subsets of projected pixels which are on the two half lines that form the triangle, denoted as $P(l^h)$ and $P(\tilde{l}^h)$, as opposed to the entire sets of projected pixels $P(l)$ and $P(\tilde{l})$.

4.2.3 Evaluating the Fitted Triangle

In order to evaluate the quality of a fitted triangle, we first introduce a *Continuity Ratio* score to evaluate the quality of a fitted side. Here we use one half line l^h as an example. The way to calculate continuity ratio for the other half line \tilde{l}^h is exactly the same. Given any point X lying on the half line l^h , we can construct a potential side OX connecting the intersection point $O = (x_o, y_o)$ and X . We further compute the number of pixels projected onto OX divided by the length of OX as $\frac{P(l^h) \cap OX}{|OX|}$. Then, the *Continuity Ratio* of l^h is defined as the ratio of the best fitted side on l^h :

$$C(l^h) = \max_{X \in P(l^h)} \frac{P(l^h) \cap OX}{|OX|}. \quad (2)$$

Finally, the continuity ratio for the entire triangle constructed by the two half lines l^h and \tilde{l}^h is

$$C(l^h, \tilde{l}^h) = C(l^h) \times C(\tilde{l}^h). \quad (3)$$

In addition to the continuity ratio, we define another *Total Ratio* score which is calculated as the area of the triangle divided by the area of the entire picture. Intuitively, the *Continuity Ratio* describes how well the extracted line segments fit the given side. Meanwhile, the *Total Ratio* represents the significance of a fitted triangle in terms of sizes. As a bigger triangle can be more easily recognized and has more impact on composition of the entire image, we only keep the triangles whose *Continuity Ratio* and *Total Ratio* scores are both above certain thresholds.

5 Experiments

5.1 Image Segmentation for Natural/Urban Scenes

In this section, we compare the performance of our method with the state-of-the-art image segmentation method, *gPb-owt-ucm* (Arbelaez et al., 2011). For this experiment, we assume known dominant vanishing point locations. We emphasize that our goal here is not to

compete with that work as a generic image segmentation algorithm, but to demonstrate that information about the vanishing point (i.e., the geometric cue), if properly harnessed, can empower us to get better segmentation results.

To quantitatively evaluate the methods, we use three popular metrics to compare the result obtained by each algorithm with the manually-labeled segmentation: Rand index (RI), variation of information (VOI) and segmentation covering (SC). First, the RI metric measures the probability that an arbitrary pair of pixels have the same label in both partitions or have different labels in both partitions. The range of RI metric is $[0, 1]$, higher values indicating greater similarity between two partitions. Second, the VOI metric measures the average condition entropy of two clustering results, which essentially measures the extent to which one clustering can explain the other. The VOI metric is non-negative, with lower values indicating greater similarity. Finally, the SC metric measures the overlap between the region pairs in two partitions. The range of SC metric is $[0, 1]$, higher values indicating greater similarity. We refer interested readers to (Arbelaez et al., 2011) for more details about these metrics.

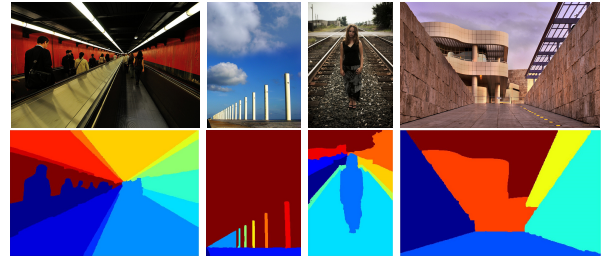


Fig. 13 Example test images with the manually labeled segmentation maps.

For this experiment, we manually labeled 200 images downloaded from *flickr.com*. These images cover a variety of indoor and outdoor scenes and each has a dominant vanishing point. During the labeling process, our focus is on identifying all the regions that differ from their neighbors either in their geometric structures or photometric properties. We show some images with the hand-labeled segmentation maps in Figure 13. Note that, ideally this process should be done by someone unrelated to the work and even by multiple people. But since this particular segmentation task is relatively well-defined, the level of subjectivity or inter-rater variation is not expected to be high. Also, the process is labor intensive as it traces region boundaries. Through doing the task ourselves, we ensure quality of the segmentation maps. The manually-labelled segmentation maps will be made available so researchers can examine for correctness and experiment with them.

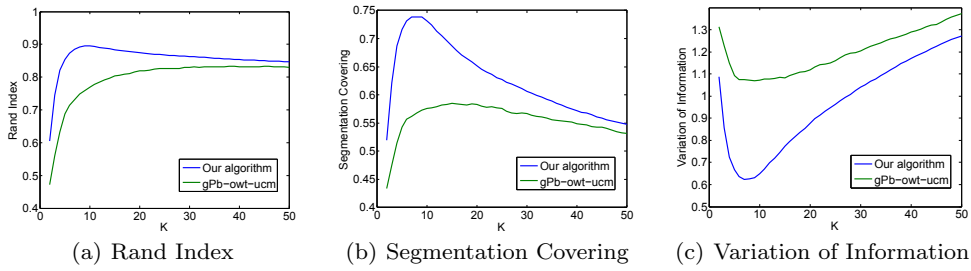


Fig. 14 Segmentation benchmarks. K is the number of regions.

Figure 14 shows the benchmark results of both methods. Our method significantly outperforms *gPb-owt-ucm* on all metrics. This is consistent with the example results in Figures 6 and 7, suggesting that our method is advantageous in segmenting the geometric structures in the scene.

5.2 Vanishing Point Detection

Next, we compare our vanishing point detection method with two state-of-the-art methods proposed by Tardif (2009) and Tretiak et al. (2012), respectively. As discussed earlier, both methods rely on the line segments to generate vanishing point candidates. Then, a non-iterative scheme similar to the popular RANSAC technique is developed by Tardif (2009) to group the line segments into several clusters, each corresponding to one vanishing point. Using the vanishing points detected by Tardif (2009) as an initialization, Tretiak et al. (2012) further propose a non-linear optimization framework to jointly refine the extracted line segments and vanishing points.

In this experiment, we use 400 images downloaded from [flickr.com](https://www.flickr.com) whose dominant vanishing points lie within the image frame. All images are scaled to size 500×330 or 330×500 . To make the comparison fair, for Tardif (2009) and Tretiak et al. (2012) we only keep the vanishing point with the largest support set among all hypotheses that also lie within the image frame. We consider a detection successful if the distance between the detected vanishing point and the manually labeled ground truth is smaller than certain threshold t , and plot the success rates of all methods w.r.t. the threshold t in Figure 15. Our method outperforms existing methods as long as the threshold is not too small ($t \geq 10$ pixels), justifying its effectiveness for detecting the dominant vanishing point in arbitrary images. When t is small, our method does not perform well because its precision in locating the vanishing point is limited by the size of the grid mesh. Nevertheless, this issue can be alleviated using a denser grid mesh at the cost of more

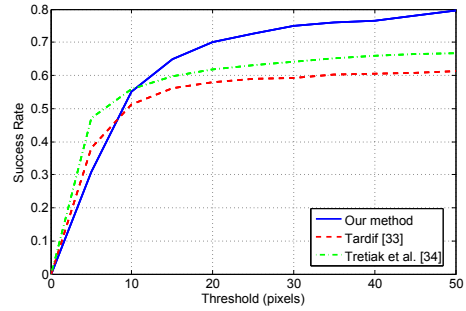


Fig. 15 Comparison of vanishing point detection algorithms.

computational time. Also, we note that while the joint optimization scheme proposed by Tretiak et al. (2012) can recover weak line segments and vanishing points for urban scenes, its improvement over Tardif (2009) is quite small in our case.

5.3 Triangle Detection in Portrait Images

In order to evaluate the performance of our triangle detection method for portrait images, we build a dataset by collecting 4,451 professional studio portrait photos from Flickr.

Figure 16 presents some detected triangles with different continuity ratios. As one can see, triangles with high continuity ratios are more easily recognized than those with low continuity ratios. However, they do not necessarily outperform those of low continuity ratios in conveying useful compositional information about the photo. For instance, Figure 16(j) has a much lower continuity ratio than Figure 16(m) but it conveys a more interesting compositional skill. From the detected triangle, we can notice that the model slightly tilts her head to align with her left arm which constructs a beautiful triangle with her hair. It is very common that multiple different triangles are embedded in one image, and some of them can be easily overlooked by amateurs. Here, our goal is to identify all potential triangles from images, which enables amateur photographers to better learn compositional techniques.

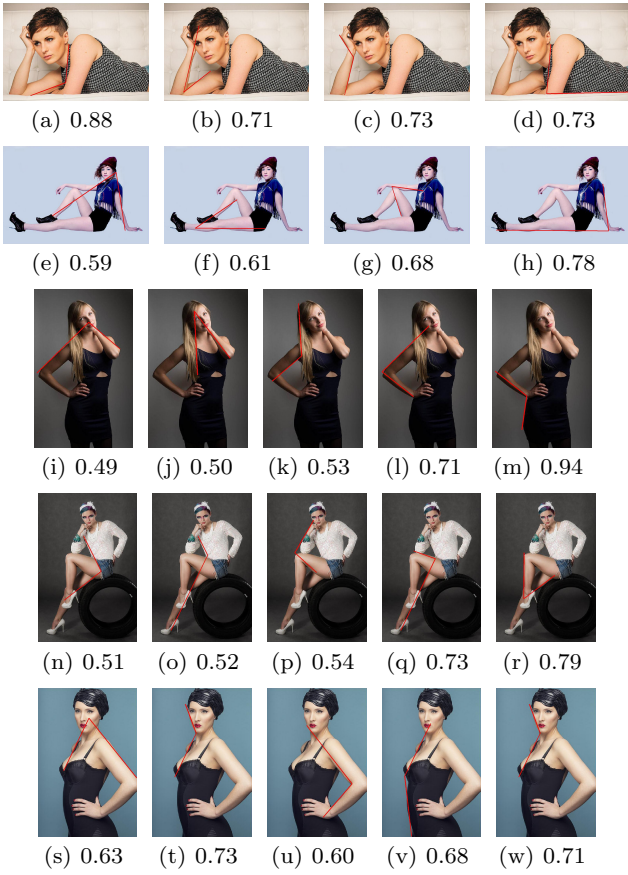


Fig. 16 Detected triangles with different *Continuity Ratios*. The red lines indicate two sides of the detected triangle.

More detected results can be found in Figure 17. It demonstrates that our triangle detection system can clearly identify triangles in portrait photographs despite the existence of noises such as human hair, props, and shadows/patterns/folds on shirts, etc. Moreover, triangles involving multiple human subjects can be accurately detected as well (*e.g.*, the fourth picture in the first row and the third picture in the second row). More interestingly, our system is able to locate triangles that may not be easy to identify by human eyes. For instance, considering the first photo in the last row, it is quite easy to find the triangle containing the girl’s two arms. However, we often overlook another triangle which is constructed with one arm of the girl and the edge of her lower jaw. Another example is the fifth picture in the last row. The girl puts her arm on her dress in a deliberately designed pose so that it extends the boundary of her dress and forms a big triangle together with her long hair. Both examples indicate that professional photographers usually design delicate poses for the subjects in order to achieve quality photo composition. However, choosing an interesting pose requires

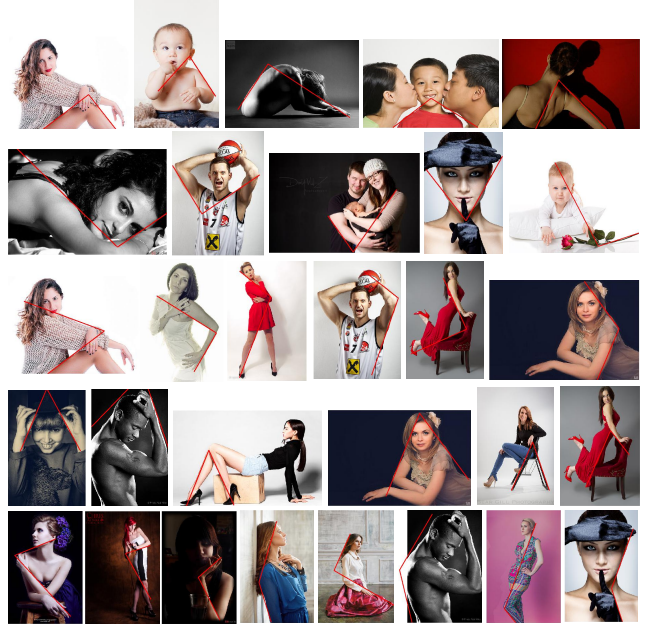


Fig. 17 Examples of detected triangles in portrait photographs.

much experience and artistic inspiration. The triangles detected by our system can help amateurs gain deeper understanding and inspirations from high-quality photography works.

Quantitative Evaluation. To further study the effectiveness of our method, we involved an experienced professional portrait photographer, who has studied arts and architecture and has operated a professional portrait studio for over 30 years. Using the online annotation tool LabelMe¹, we asked the photographer to manually annotate triangles that indicate interesting pose or composition in the images we collected. A total of 173 images were annotated by the photographer. We show some example triangles in Figure 18, which illustrate a wide variety of numbers, sizes and orientations of triangles used in the portrait photography. As the professional annotations can be valuable to computer vision community in studying portrait composition, we will make the dataset and any future extensions freely available to researchers.

We compare the triangles detected by our method with the manual annotations. In this experiment, we only consider triangles whose continuity ratio and total ratio are above 0.1. Let (A, B, C) denote the set of vertices of a ground truth triangle, we consider a candidate triangle (A', B', C') a matching triangle if

$$\frac{|AA'| + |BB'| + |CC'|}{|AB| + |BC| + |CA|} \leq \delta, \quad (4)$$

¹ <http://labelme.csail.mit.edu/Release3.0/>

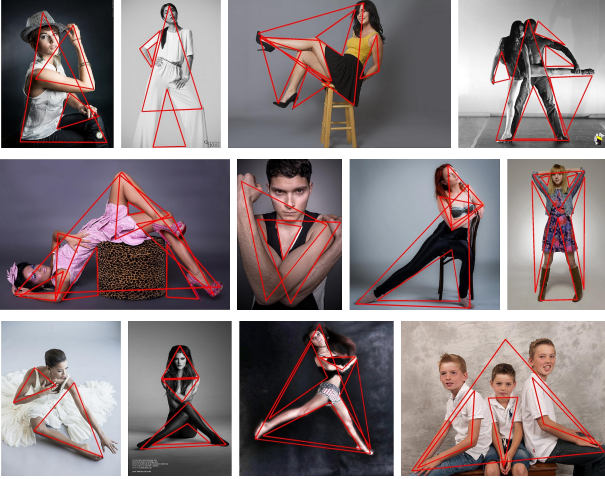


Fig. 18 Example portrait photos with triangles labeled by an experienced professional photographer.

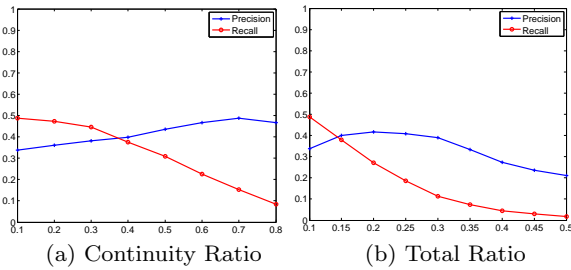


Fig. 19 Quantitative evaluation of triangle detection in portraits. The plots show the precision and recall of our method as a function of (a) continuity ratio and (b) total ratio.

where $|AB|$ is the length of the line segment connecting A and B , and δ is a threshold. We fix $\delta = 0.3$ in this paper.

In Figure 19, we report the *precision* and *recall* of our method as a function of the continuity ratio and total ratio. Specifically, let G denote the set of ground truth triangles annotated by the professional photographer, and Q denote the set of triangles detected by our algorithm under a particular experiment setting, the precision and recall are defined as follows:

$$Precision = \frac{|G \cap Q|}{|Q|}, \quad Recall = \frac{|G \cap Q|}{|G|}. \quad (5)$$

As shown in Figure 19(a), the precision of our method increases as the continuity ratio increases. When the continuity ratio is high, about half of the triangles detected are true positives. Meanwhile, among all manually annotated triangles, up to about half of them can be detected by our method (i.e., when continuity ratio is 0.1). In Figure 20 (first row) we show some triangles missed by our algorithm. As one can see, for many of such triangles, there is a lack of explicit edges or contours in the image. For example, in the first image of

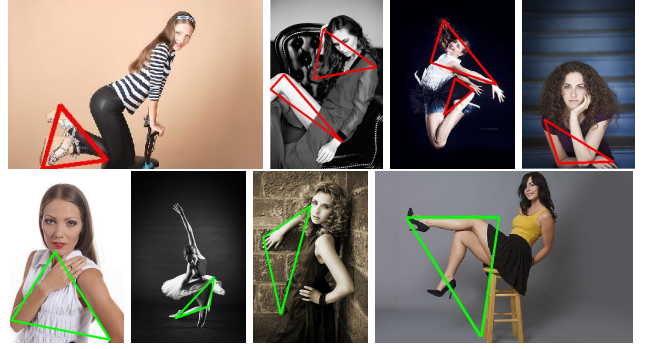


Fig. 20 Comparison of triangles detected by our method with professional annotations. **First row:** Triangles annotated by the professional photographer but missed by our algorithm. **Second row:** Some interesting triangles detected by our algorithm, but not labeled by the photographer.

Figure 20, the triangle is formed by the pair of shoes and the girl’s knees, instead of explicit edges. Such triangles are often known as *implicit triangles* in photography. It will be interesting future work to look for them using machine vision. In addition, Figure 19(b) shows that our method achieves the best precision when the total ratio is about 0.2. This may suggest the most commonly seen triangle sizes in portrait photography.

Finally, we examine the triangles that are detected by our algorithm but not labeled by the professional photographer. Figure 20 (second row) shows some interesting cases where we think the triangles are actually meaningful. These examples suggest that even experienced photographers may occasionally overlook certain elements, and our algorithm could potentially provide them with an alternative interpretation of the photo.

6 Application in On-Site Composition Feedback

The proposed triangle technique detection methods capture rich information about the composition of photographs. They can be integrated with various composition-driven applications. Here, we discuss an application that aims at providing amateur users with on-site feedback about the composition of their photos.

6.1 Natural/Urban Scene Photography

For natural/urban scene photography, given a photo taken by a user, we propose to find photos with similar compositions in a collection of photos taken by experienced or accomplished photographers. These photos are rendered as feedback to the user. The user can then examine these exemplar photos and consider re-composing his/her own photo accordingly, while the user remains on-site. Yao et al. (2012) pioneered this

direction, but the types of composition studied there are limited to a few categories which are pre-defined based on simple 2D rules.

In this paper, we take a completely different approach and develop a similarity measure to compare the composition of two images based on their geometric image segmentation maps. Our observation is that, experienced photographers often are able to achieve different compositions by first placing the dominant vanishing point at different image locations, before choosing how the main structures of the scene are related to it in the captured image. In addition, while the difference in the dominant vanishing point locations can be simply computed as the Euclidean distance between them, our geometric segmentation result offers a natural representation of the arrangement of structures with respect to the dominant vanishing point. Specifically, given two images I_i and I_j , let P_i and P_j be the locations of dominant vanishing points and S_i and S_j be the segmentation results generated by our method for these two images, respectively, we define the similarity measure as:²

$$D(I_i, I_j) = F(S_i, S_j) + \alpha \|P_i - P_j\|, \quad (6)$$

where $F(S_i, S_j)$ is a metric to compare two segmentation maps. We adopt the Rand index (Rand, 1971) for its effectiveness. In addition, α controls the relative impact of the two terms in the equation. We empirically set $\alpha = 0.5$.

To obtain a dataset of photos that make good use of the perspective effect, we collect 3,728 images from [flickr.com](http://www.flickr.com) by querying the keyword “vanishing point”. When collecting the photos, we use the sorting criterion of “interestingness” provided by [flickr.com](http://www.flickr.com), so that the retrieved photos are likely to be well composed and taken by experienced or accomplished photographers. Each photo is then scaled to size 500×330 or 330×500 . To evaluate the effectiveness of our similarity measure (Eq. (6)), we manually label the dominant vanishing point and then apply our geometric image segmentation algorithm (with the proposed distance measure $W(e_{ij})$ and the stopping criteria $\delta = 0.55$) to obtain a segmentation for each image.

In Figure 21, we show the retrieved images for various query images. The results clearly show that the proposed measure is not only able to find images with similar dominant vanishing point locations, but also effectively captures how each region in the image is related to the vanishing point.

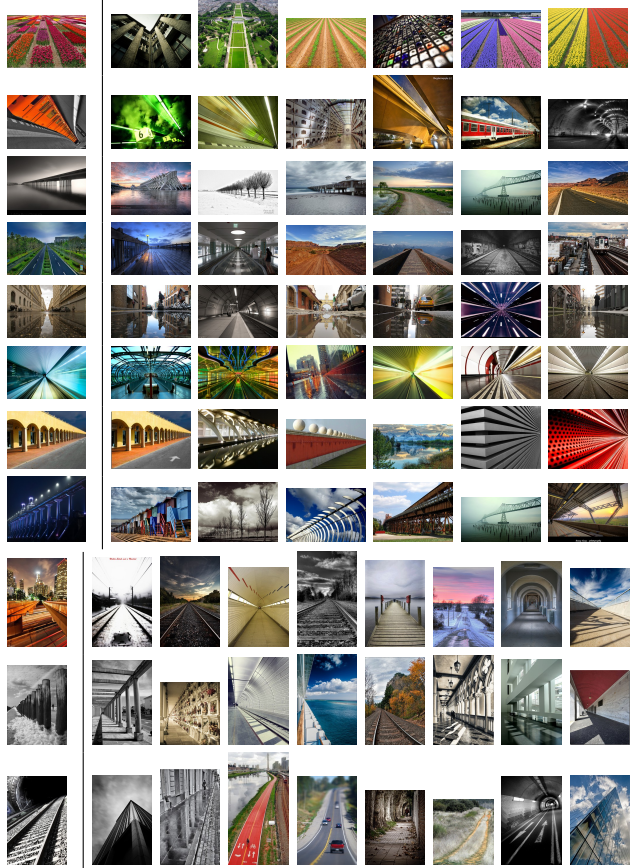


Fig. 21 Composition-sensitive image retrieval results. Each row shows a query image (first image from the left) and the top-6 or top-8 ranked images retrieved.

6.1.1 Comparison to Existing Retrieval Systems

We also compare our composition-sensitive image retrieval system to the following recent retrieval pipelines:

HOG: We represent each image with a rigid grid-like HOG feature \mathbf{x}_i (Dalal and Triggs, 2005; Felzenszwalb et al., 2010). In order to limit the dimensionality of HOG features to roughly $5K$, we resize the images to 150×100 or 100×150 , and use a cell size of 8 pixels. As suggested by Shrivastava et al. (2011), we further normalize the feature vector by subtracting its mean: $\mathbf{x}_i = \mathbf{x}_i - \text{mean}(\mathbf{x}_i)$, and use the cosine distance to measure the similarity of two vectors.

VLAD: The vector of locally aggregated descriptors (VLAD) is a feature coding and pooling method (Jegou et al., 2010; Arandjelovic and Zisserman, 2013). It encodes a set of local feature descriptors (e.g., SIFT features) extracted from an image using a dictionary built using a clustering method such as GMM or K-means. In this paper, we use the code provided on the authors’ website³ with pre-trained dictionary to extract

² Here, we assume the two images have the same size, after rescaling.

³ <http://people.rennes.inria.fr/Herve.Jegou/software.html>

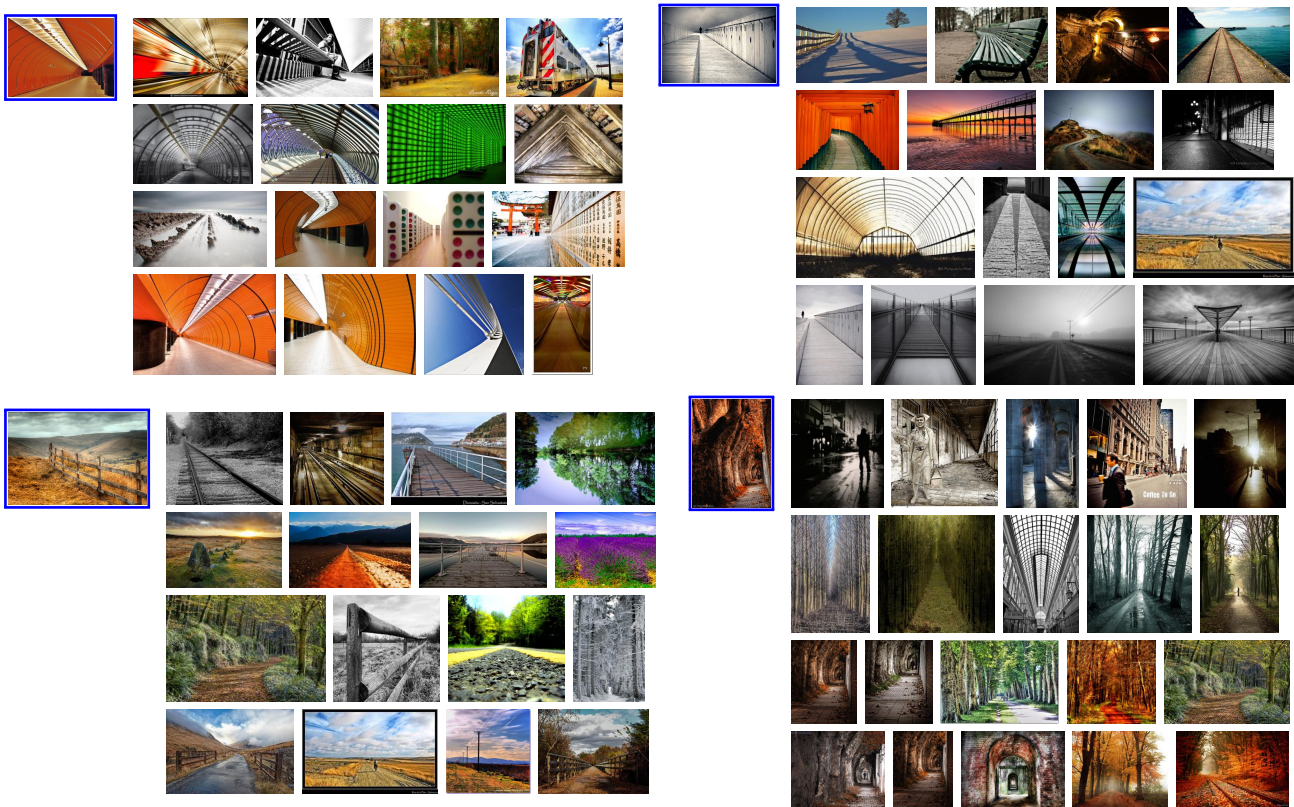


Fig. 22 Comparison of four retrieval systems for images of similar composition. For a query image, we show the top-4 or top-5 images retrieved by four different systems, where each row corresponds to one system. The four rows are ordered as follows: **First row:** our system. **Second row:** HOG. **Third row:** VLAD. **Fourth row:** CNN.

the VLAD descriptor for each image, and compare different descriptors using the ℓ_2 distance.

CNN: Generic descriptors extracted from the convolutional neural networks (CNNs) have been shown to be very powerful in tackling a diverse range of computer vision problems, including image retrieval (Razavian et al., 2014). In this paper, we use the publicly available code and model⁴ by Chatfield et al. (2014), which were developed to perform classification in the ImageNet ILSVRC challenge data, and represent each image using the ℓ_2 -normalized output of the second fully connected layer (full7 of (Chatfield et al., 2014)). The feature similarity is measured by the cosine distance.

Figure 22 shows the top-ranked images retrieved by all four systems for some example query images. As can be seen, the images retrieved by our system is more compositionally relevant in terms of the use of perspective effect than other systems. Among the three existing systems, HOG is shown to be more sensitive to the image composition, as it is based on the local image gradients. Meanwhile, VLAD and CNN features are known

to perform well in capturing the semantics of a scene (i.e., scene types and objects). While both methods indeed retrieved more relevant images semantically, they are not sensitive to the image composition.

Quantitative Evaluation. Unlike traditional image retrieval tasks, currently there is no dataset with ground truth composition labels available. In order to quantitatively evaluate the performance of our system, we have conducted a user study that allows participants to manually rank the performance of the four systems based on *their ability to retrieve compositionally similar images*.

In this study, a collection of 200 query images were randomly selected to form the dataset for the comparison study. At an online website, each participant is provided with a subset of 15 randomly selected query images. For each query, we show the participant the top-8 images retrieved by all four systems. Then, we ask the users to rank the performance of the four systems without providing them with any information about the four systems. To avoid any biases, we also randomly shuffled the order in which the results of the four systems are presented on each page.

⁴ <http://www.vlfeat.org/matconvnet/pretrained/>

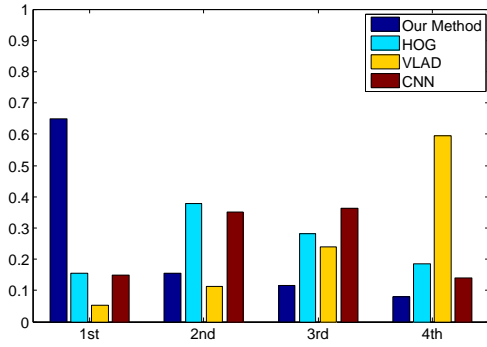


Fig. 23 Quantitative evaluation of retrieval systems. Percentages of user selections are shown. Our system is ranked the first by the participants 65.6% of the time.

To complete this study, we recruited 20 participants, mostly graduate students at Penn State with some basic photography knowledge. Figure 23 shows the overall percentage of times that each system is ranked the first, the second, the third, and the fourth, respectively. Our system is ranked the first 65.6% of the time, which is significantly higher than the other systems. Among the other three systems, HOG outperforms VLAD and CNN, thanks to its sensitivity to local image gradients, which is correlated with the vanishing direction of an image. Interestingly, while CNN is trained to capture scene semantics, it is voted first for 15% of the time. This suggests a link between semantics (*i.e.*, what is in the image) and image composition (*i.e.*, how an image is taken).

6.2 Portrait Photography

Our triangle technique detection method can potentially help amateur photographers design interesting poses when taking portraits. It is often difficult for untrained photographers to naturally embed triangles in compositions in order to form striking portraits. Certain professional users, *e.g.*, magazine editors, can also benefit from the triangle detection method. When they are selecting portraits to fill a specific page layout, the shapes and orientations of embedded triangles within these photos can be critical to the overall page composition. Therefore, we develop a portrait retrieval system that can take a *triangle sketch* as a query to help users find images containing a targeted triangular configuration. Such tools can be especially useful when a large collection of portraits are available to choose from.

First, users need to provide a sketch indicating the shape and the orientation of the triangular configuration that they desire to have in the photos. In portrait photos, the third side of a triangle is often missing in most photos. Thus, the sketch query provided by users

is basically an angle with two sides. Given the sketch angle, we compute the *orientations* of the two sides as well as the *opening direction* of the angle. Specifically, the orientation of a side is defined to be the angle between its extended straight line and the positive x -axis. An angle has four possible opening directions: upward, downward, leftward, or rightward. The two properties narrow down the searching space to a specific type of triangles.

Recall that our triangle detection algorithm has two steps: line segment detection and fitting triangles. All line segments detected from the first step are taken as candidate triangle sides during the fitting stage. In the retrieval system, we construct two candidate sets containing line segments with similar orientations as the two sides of the sketched angle. Two sides can then be randomly selected from the two candidate sets and the combination of them generates four angles with four different opening directions (Figure 12(d)). Only the angle that has the same opening direction as the sketched angle will be taken into consideration during the fitting process. Such a sketch-based triangle retrieval system not only assists users in searching for desired types of triangles but also reduces the searching time significantly for large photo collections.

6.2.1 User Study

We conducted a human subject study to verify the effectiveness of the retrieved triangles in conveying valuable information about composition to amateurs. In this study, we selected 20 groups of representative queries which covered a wide range of angles in terms of magnitude and orientation. We only consider angles in the range of $[45^\circ, 135^\circ]$ because angles that are either too large or too small are often not perceived as interesting ones by humans. Each of the 20 groups of queries takes a distinct combination of orientations for two straight lines. Twenty line combinations $\langle l_1, l_2 \rangle$ are selected in our experiment such that the angle between l_1 and l_2 falls in the closed range of $[45^\circ, 135^\circ]$ and the angle between l_1/l_2 and positive x -axis falls in $\{0^\circ, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ, 112.5^\circ, 135^\circ\}$. Moreover, one combination of two straight lines generates four possible angles which differ in terms of their opening directions. Therefore, we use a total of 80 queries to retrieve triangles from 4,451 photos where each photo may contain many distinct triangles. For each query, we rank the results based on their continuity ratios. Higher continuity ratios represent higher quality of fitting and thus imply more accurate retrieved results.

We recruited 20 participants to this study, mostly graduate students at Penn State with some basic pho-

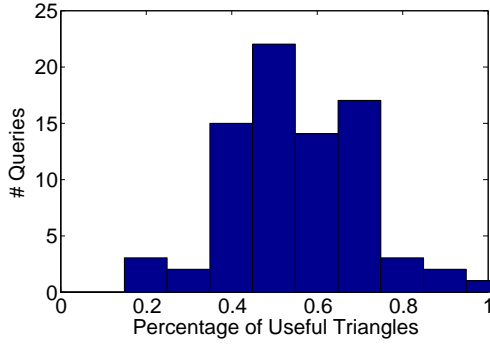


Fig. 24 The performance of the retrieval system for portraits.

tography knowledge. At an online website, each participant is provided with a subset of 15 randomly selected queries. For each query, we show the participant the top-20 triangles retrieved by our system. For each retrieved triangle, we ask the participant to assess whether it is “useful”, *i.e.*, whether it indicates interesting pose or composition in the image, and help the participant understand the use of triangles in the photo.

Figure 24 shows the histogram of overall percentages of “useful” triangles for all 80 queries. For most queries, between 40% and 80% of the retrieved results are considered by the participants as providing useful information and guidance on the portrait composition. Overall, 53.8% of the retrieved triangles are regarded as “useful” by the users.

In Figure 25, we provide examples of both “useful” and “unuseful” triangles retrieved by our system. The first column contains twelve queries among which eight return high percentages of “useful” triangles and four return low percentage of “useful” triangles. From the retrieved examples, it can be seen that professional photographers are skillful at using all kinds of objects, such as arm, leg, shoulder, chair, wall, ground, hair, apparel, or even shadow, to construct triangles. Interestingly, very often a slight adjustment of a pose can form beautiful triangles which make the entire composition aesthetically appealing. For instance, the girl in row 3, column 3 slightly turns her head towards left to perfectly align with her shoulder. For the same reason, the girl in row 1, column 4 lifts up her head a little bit.

In addition, our system also retrieves many “unuseful” triangles for some queries. As shown in the last four rows of Figure 25, many of these triangles are right triangles. In fact, it is known that professional photographers often avoid 90° body angles because they often look unnatural and strained (Valenzuela, 2012). This may partly explain why we are unable to retrieve more “useful” triangles in these cases.

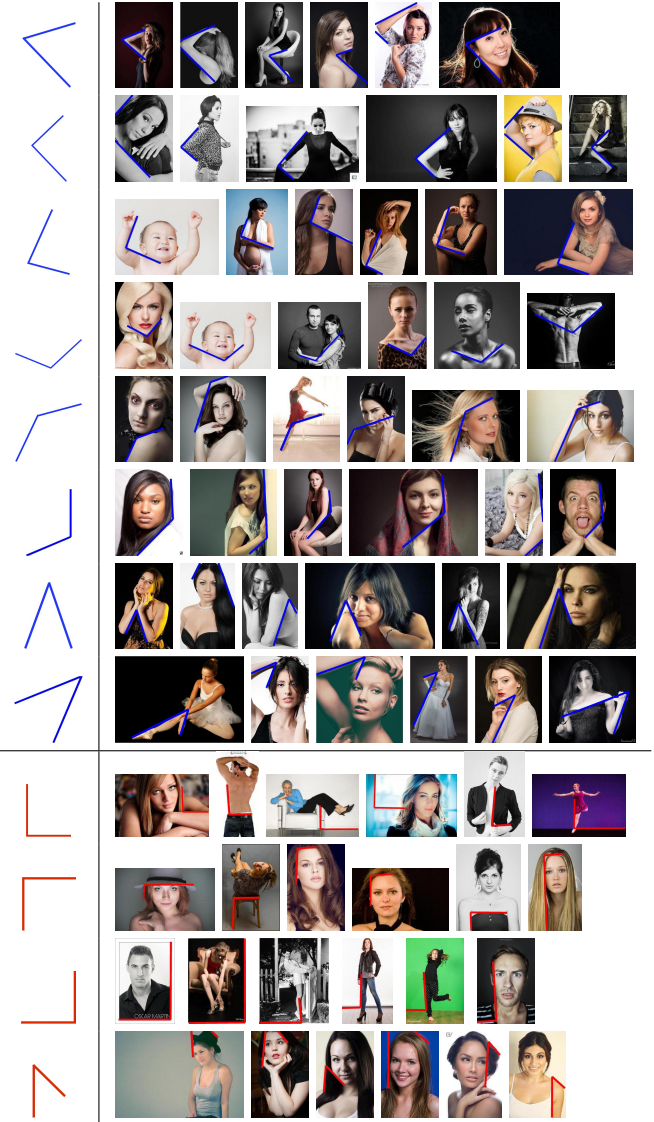


Fig. 25 Examples of retrieved portraits. Each row shows results from a query. Red examples are considered “unuseful” in our evaluation.

7 Conclusions and Future Work

This paper proposes a system that detects the usage of triangle photo composition techniques in both natural/urban scene and portrait/people photography. We show preliminary evidence through human subject studies that such systems can potentially help consumer photographers learn about professional composition of photographs. The two broad categories that we have chosen cover most major consumer photography genres. Additionally, we show that line-based methods are effective for both categories, despite the fact that many photos have no clearly-defined straight lines. For photographs of natural/urban scene with strong perspec-

tive effect, we have demonstrated a new method for modeling visual composition through analyzing the perspective effects and segmenting the image based on photometric and geometric cues. The method can effectively detect the dominant vanishing point from an arbitrary scene. For portrait photographs, we extract a set of candidate line segments from a photo and then successfully fit a triangle to these segments despite a large proportion of outliers. The fitted result accurately identifies the presence of triangles in photographs. Among a variety of potential applications, we have illustrated how our techniques can provide on-site feedback to photographers.

Our work opens up several future directions. First, we plan to investigate the relationships between triangles and other visual elements and design principles. For example, one challenge in composition recognition for real-world photos is the presence of large foreground objects. They typically correspond to regions which are not associated with any vanishing point in the image. In addition, some photos may solely focus on the objects (e.g., a flower) and do not possess a well-defined perspective geometry. We will analyze the composition of these images by first separating the foreground objects from the background. We note that, while our analysis of the perspective geometry provides valuable information about the 3D space, many popular composition rules studied in early work, such as the simplicity of the scene, golden ratio, rule of thirds, and visual balance have focused on the arrangement of objects in the 2D image plane. We believe that combining the strength of both approaches will enable us to obtain a deeper understanding of the composition of these images.

Beyond image composition, the relationship between triangles and the aesthetic quality of compositions can be further studied. For example, photographers may use different composition techniques in different situations. How to assess the relevance of perspective in a natural/urban scene photo? Also, for portraits, how do the number, sizes, shapes, and orientations of triangles influence the aesthetics of photo composition? Answering such questions can help amateur photographers learn more specific photography techniques.

Finally, apart from providing on-site feedback to photographers, our method can also be implemented as a component in large-scale image retrieval engines in the cloud. When a query results in a large number of images that have similar levels of visual similarity or aesthetic quality, the query results can be structured as a tree with levels of refinement in terms of composition by grouping the images using a hierarchical clustering scheme.

Acknowledgements Chuck S. Fong of the Studio 2 Photography provided ground truth annotations for the professional portrait dataset used in our evaluation. Edward Chen and Sahil Mishra assisted in developing the data-collection system for the human subject studies. We thank the participants in the studies for their assistance. The authors would also like to acknowledge the comments and suggestions from the reviewers and the guest editors.

References

- Arandjelovic R, Zisserman A (2013) All about VLAD. *In: 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp 1578–1585
- Arbelaez P (2006) Boundary extraction in natural images using ultrametric contour maps. *In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*
- Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(5):898–916
- Barinova O, Konushin V, Yakubenko A, Lee K, Lim H, Konushin A (2008) Fast automatic single-view 3-d reconstruction of urban scenes. *In: Proceedings of the European Conference on Computer Vision: Part II*, pp 100–113
- Bhattacharya S, Sukthankar R, Shah M (2010) A framework for photo-quality assessment and enhancement based on visual aesthetics. *In: Proceedings of the ACM International Conference on Multimedia*, pp 271–280
- Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. *In: British Machine Vision Conference*
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. *In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp 886–893
- Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. *In: Proceedings of the European Conference on Computer Vision-Volume Part III*, pp 288–301
- Dhar S, Ordonez V, Berg TL (2011) High level describable attributes for predicting aesthetics and interestingness. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 1657–1664
- Fang C, Lin Z, Mech R, Shen X (2014) Automatic image cropping using visual composition, boundary simplicity and content preservation models. *In: Proceedings of the ACM International Conference on Multimedia*, pp 1105–1108
- Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. *International Journal of Computer Vision* 59(2):167–181
- Felzenszwalb PF, Girshick RB, McAllester DA, Ramanan D (2010) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9):1627–1645
- von Gioi RG, Jakubowicz J, Morel J, Randall G (2010) LSD: A fast line segment detector with a false detection control. *IEEE Trans Pattern Anal Mach Intell* 32(4):722–732
- Gould S, Fulton R, Koller D (2009) Decomposing a scene into geometric and semantically consistent regions. *In: Proceedings of the IEEE International Conference on Computer Vision*, pp 1–8
- Gupta A, Efros AA, Hebert M (2010) Blocks world revisited: Image understanding using qualitative geometry and

- mechanics. In: *Proceedings of the European Conference on Computer Vision: Part IV*, pp 482–496
- Han F, Zhu SC (2009) Bottom-up/top-down image parsing by attribute graph grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1):59–73
- Hedau V, Hoiem D, Forsyth DA (2009) Recovering the spatial layout of cluttered rooms. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1849–1856
- Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: Using appearance models and context based on room geometry. In: *Proceedings of the European Conference on Computer vision: Part VI*, pp 224–237
- Hoiem D, Efros AA, Hebert M (2005) Automatic photo pop-up. *ACM Transactions on Graphics* 24(3):577–584
- Hoiem D, Efros AA, Hebert M (2007) Recovering surface layout from an image. *International Journal of Computer Vision* 75(1):151–172
- Jegou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, pp 3304–3311
- Jin X, Zhao M, Chen X, Zhao Q, Zhu SC (2010) Learning artistic lighting template from portrait photographs. In: *Proceedings of the European Conference on Computer vision: Part IV*, pp 101–114
- Kong H, Audibert JY, Ponce J (2009) Vanishing point detection for road detection. In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp 96–103
- Kosecká J, Zhang W (2002) Video compass. In: *Proceedings of the European Conference on Computer Vision-Part IV*, pp 476–490
- Lauer DA, Pentak S (2011) *Design Basics*. Cengage Learning
- Lee DC, Hebert M, Kanade T (2009) Geometric reasoning for single image structure recovery. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 2136–2143
- Li J (2011) Agglomerative connectivity constrained clustering for image segmentation. *Statistical Analysis and Data Mining* 4(1):84–99
- Liu L, Chen R, Wolf L, Cohen-Or D (2010) Optimizing photo composition. *Computer Graphics Forum* 29(2):469–478
- Luo Y, Tang X (2008) Photo and video quality evaluation: Focusing on the subject. In: *Proceedings of the European Conference on Computer Vision: Part III*, pp 386–399
- Marchesotti L, Perronnin F, Larlus D, Csurka G (2011) Assessing the aesthetic quality of photographs using generic image descriptors. In: *Proceedings of IEEE International Conference on Computer Vision*, pp 1784–1791
- Mobahi H, Rao S, Yang AY, Sastry SS, Ma Y (2011) Segmentation of natural images by texture and boundary compression. *International Journal of Computer Vision* 95(1):86–98
- Nedovic V, Smeulders AWM, Redert A, Geusebroek JM (2010) Stages as models of scene geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9):1673–1687
- Obrador P, Schmidt-Hackenberg L, Oliver N (2010) The role of image composition in image aesthetics. In: *Proceedings of the IEEE International Conference on Image Processing*, pp 3185–3188
- Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846–850
- Rasmussen C (2004) Grouping dominant orientations for ill-structured road following. In: *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition*, pp 470–477
- Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: An astounding baseline for recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, pp 512–519
- Russell BC, Efros AA, Sivic J, Freeman B, Zisserman A (2009) Segmenting scenes by matching image composites. In: *Advances in Neural Information Processing Systems*, pp 1580–1588
- Saxena A, Sun M, Ng AY (2009) Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(5):824–840
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905
- Shrivastava A, Malisiewicz T, Gupta A, Efros AA (2011) Data-driven visual similarity for cross-domain image matching. *ACM Trans Graph* 30(6):154
- Su HH, Chen TW, Kao CC, Hsu WH, Chien SY (2011) Scenic photo quality assessment with bag of aesthetics-preserving features. In: *Proceedings of the ACM International Conference on Multimedia*, pp 1213–1216
- Tardif JP (2009) Non-iterative approach for fast and accurate vanishing point detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1250–1257
- Tretiak E, Barinova O, Kohli P, Lempitsky VS (2012) Geometric image parsing in man-made environments. *International Journal of Computer Vision* 97(3):305–321
- Valenzuela R (2012) *Picture Perfect Practice: A Self-Training Guide to Mastering the Challenges of Taking World-Class Photographs*. New Riders
- Yao L, Suryanarayan P, Qiao M, Wang JZ, Li J (2012) Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision* 96(3):353–383
- Zhang Y, Sun X, Yao H, Qin L, Huang Q (2012) Aesthetic composition representation for portrait photographing recommendation. In: *Proceedings of the IEEE International Conference on Image Processing*, pp 2753–2756
- Zhou Z, He S, Li J, Wang JZ (2015) Modeling perspective effects in photometric composition. In: *Proceedings of the ACM International Conference on Multimedia*, pp 301–310